N84-25325

MPP DISK SUBSYSTEM

GOODYEAR AEROSPACE CORPORATION
1210 MASSILLON ROAD
AKRON, OHIO 44315

MARCH, 1984

FINAL REPORT

PREPARED FOR

GODDARD SPACE FLIGHT CENTER

GREENBELT, MARYLAND 20771

PREFACE

The MPP Disk Study was intended to produce a block level design for a mass storage subsystem to be attached to the MPP. The subsystem was to have a storage capacity of 1000 MBytes initially, expandable to 5000 MBytes, and a transfer rate to the MPP stager of 10 MByte/sec., expandable to 40 MByte/sec.

The study has produced two designs: the first has a capacity of 4992 MBytes, expandable to 39936 MBytes, and a transfer rate of 25 MByte/sec, expandable to 100 MByte/sec. The second design has a capacity of 2496 MByte and a transfer rate of 10.6 MByte/sec., and is expandable via additional hardware and software to a capacity of 29952 MBytes, and a transfer rate of 84.8 MByte/sec.

Preliminary estimates place the cost of the first design at approximately $3.4 million, and the cost of the second design at approximately $900,000. The implementation schedule for the first design is 18 months, while that for the second design is 12 months.

TECHNICAL REPORT STANDARD TITLE PAGE

1. Report No.	2. Government Accession No.	3. Recipient's Catalog No.

4. Title and Subtitle	5. Report Date
MPP DISK SUBSYSTEM	26 March 1984
	6. Performing Organization Code

7. Author(s)	8. Performing Organization Report No.
W. A. Hudgins	GER-17234

9. Performing Organization Name and Address	10. Work Unit No.
Goodyear Aerospace Corporation 1210 Massillon Rd. Akron, OH 44315	11. Contract or Grant No. NA55-27613
	13. Type of Report and Period Covered FINAL 10-83 TO 3-84
12. Sponsoring Agency Name and Address J. R. Fischer, Technical Officer Goddard Space Flight Center Greenbelt, MD 20771	14. Sponsoring Agency Code

15. Supplementary Notes

16. Abstract

A disk subsystem for the Massively Parallel Processor is designed to the block diagram level. The subsystem is capable of storing 4992 megabytes of data, expandable to 39936 megabytes. The subsystem is capable of transferring data to the MPP Staging Memory at a rate of 25 megabytes/second, expandable to 100 megabytes/second.

A lower cost disk subsystem is also presented. This alternate subsystem is capable of storing 3744 megabytes with a transfer rate of 10.6 megabyte/second.

17. Key Words (Selected by Author(s))	18. Distribution Statement
Parallel Processors, Disk Subsystems MPP	

19. Security Classif. (of this report)	20. Security Classif. (of this page)	21. No. of Pages	22. Price*
None	None	49	

TABLE OF CONTENTS

1. INTRODUCTION AND SUMMARY

1.1 BACKGROUND/DESIGN GOALS

On September 27, 1983, contract #NAS5-27613 was awarded by NASA to Goodyear Aerospace Corporation (GAC), for the analysis and block-level design of a high performance disk subsystem for the MPP. This subsystem was to consist of a multi-ported interface, control module, related software and a set of commercially available disk units. The design goal for the system was an initial storage capability of 1 GByte with a transfer rate of 10 MByte/sec; expandable to a storage capability of 5 GByte with a composite transfer rate of 40 MByte/sec.

1.2 SCOPE

This final design report documents efforts and achievements by Goodyear Aerospace under the MPP Disk Subsystem Design Study. Options available for the system design are reviewed. A selected design is presented and described in detail; all pertinent performance parameters are given. The system design is modular in nature; it provides for the minimum performance specifications of the contract and expands to meet the maximum performance specifications. The design will support additional expansion to provide performance well beyond the contract specification.

Estimates are provided for cost and schedule associated with actual development of a disk system; risks associated with development are assessed.

This report also provides a top-level system design for an alternate disk system which can be developed via integration of components which have recently become available in the commercial marketplace.

1.3 ACCOMPLISHMENT

The disk subsystem, as designed, meets the contract requirements by providing the MPP with mass storage of 1248 MByte, and a transfer rate of 10.6 MByte/s., in the minimum configuration. The subsystem is expandable to a total storage capacity of 39936 MByte, and a transfer rate of 100 MByte/s.

1.4 DESIGN SUMMARY

1.4.1 DESIGN PHILOSOPHY

The mass storage subsystem was designed, at the block level, with the following design philosophy in mind:

* Reliability - The system has been designed, as much as possible, with off-the-shelf components. Data integrity is repeatedly checked. Also, since all drive controller and disk drive addresses are switch selectable, the system can be easily reconfigured in the event of the failure of one or more disk drives or drive controllers.

* Modularity - The system design makes repeated use of relatively few PCB designs. This reduces design and material cost, and also improves reliability.

* Expandability - The subsystem can be expanded to provide more disk units; the added disk drives increase bandwidth as well as system capacity. as drives are added. In its maximum configuration the subsystem has a total capacity of 39936 MByte, and supports transfer rates of up to 100 MByte/s. Also, the subsystem can be tied into an additional VAX for a more efficient production environment, and additional data ports can be added.

1.4.2 OVERVIEW AND RATIONALE

The subsystem was conceived as a way of removing the bottleneck that slows the data transfer rate into and out of the current MPP system. This bottleneck can be eliminated by:

(1) The use of high speed disks to increase raw transfer rate.

(2) Using parallelism to further increase transfer rate while simultaneously increasing storage capacity.

(3) Using data buffers to eliminate rotational latency in the drives, as well as lack of synchronism between the drives.

1.4.3 SYSTEM CONFIGURATIONS

Figure 1 shows a block diagram of the MPP disk subsystem. Subsystem storage capacity and data transfer rate are dependent on the number of disk drives and controllers employed. The minimum configuration contains one disk drive and one drive controller. The maximum configuration contains 32 disk drives and 16 drive controllers. The recommended configuration contains

four disk drives and four drive controllers.

1.4.3.1 Minimum Configuration

The minimum configuration (see Fig. 1; components in boxes with heavy outlines) consists of one disk drive and one drive controller. This subsystem would provide the MPP with 1248 MBytes of storage, and a sustained transfer rate of 10.6 MByte/s.

1.4.3.2 Maximum Configuration

The maximum configuration (see Fig. 1; components in dashed-line borders)consists of 32 disk drives, 16 drive controllers, and a secondary VAX-11/780 computer which would be responsible, primarily, for setting up data files on the disk drives, and off-loading files on which processing is complete. This configuration would provide 39936 MBytes of storage, and a transfer rate of 100 MByte/s.

1.4.3.3 Recommended Configuration

The recommended configuration of the subsystem includes four disk drives, four drive controllers, and the secondary VAX computer. This system will provide 4992 MBytes of formatted storage, and a transfer rate to the stager of 25 MByte/sec. This configuration will provide the user with a higher I/O bandwidth than the minimum system, will not load down the host VAX when loading or unloading data sets, and provides the flexibility to expand the system from 4992 MByte to 9984 MByte merely by adding drives to existing strings. This configuration provides significant performance advantages over the minimum system at a relatively small increase in price.

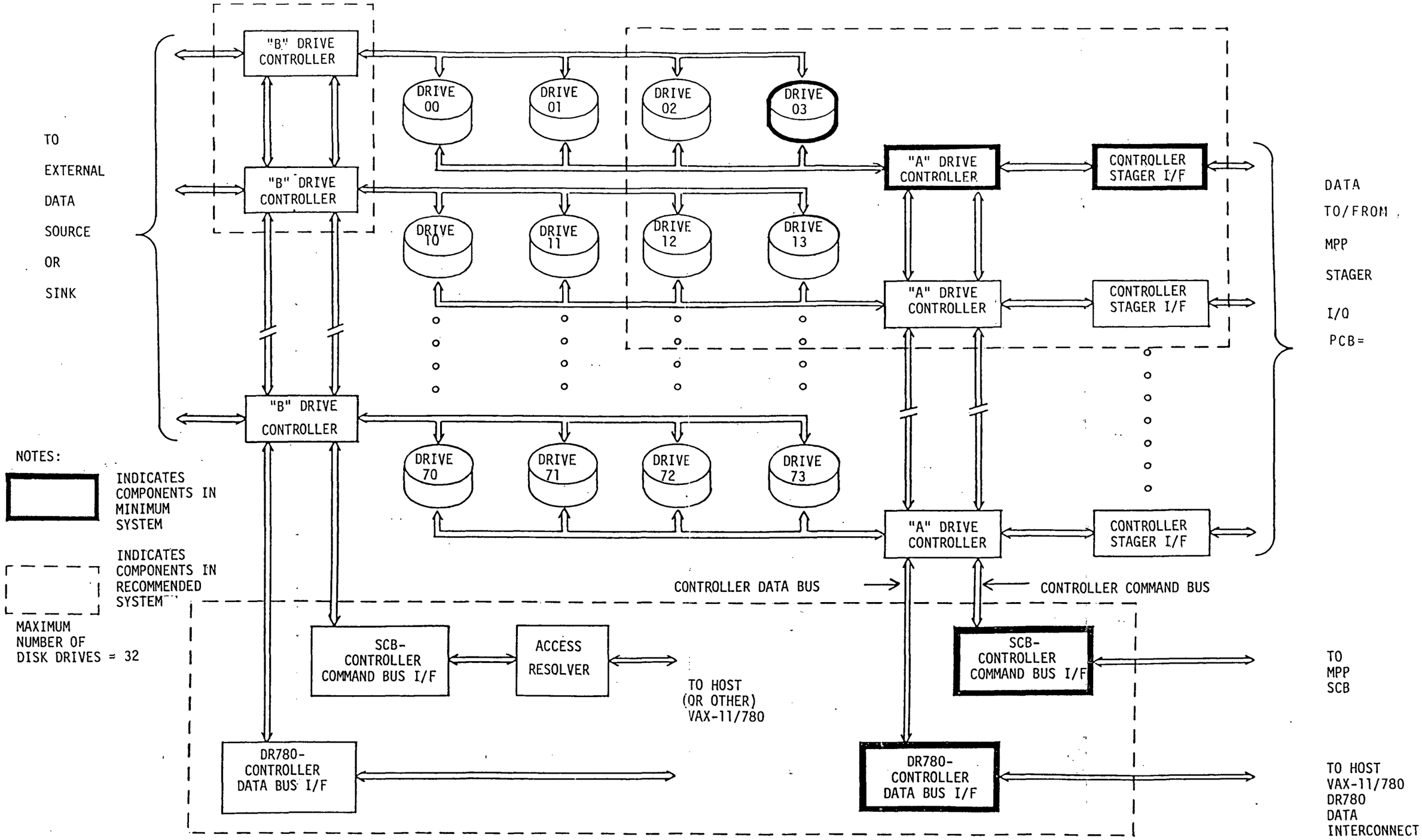


FIGURE 1: DISK SUBSYSTEM BLOCK DIAGRAM

1.4.3.4 Performance

Table 1 shows the relationship between the number of disk drives in the subsystem, the number of drive controllers, and the resulting subsystem capacity and data transfer rate. The number of controllers shown indicates only the controllers connected to the MPP side of the system, and not the "B" side drive controllers connected to the (optional) secondary VAX computer.

TABLE 1: SUBSYSTEM CAPACITY AND TRANSFER RATE MATRIX

NUMBER OF "A" CONTROLLERS	DRIVES/CONT.	TOTAL DRIVES	CAPACITY (MBYTE)	DATA RATE (MBYTE/S)
1 (MIN)	1	1	1248	10.6
	2	2	2496	12.5
	3	3	3744	12.5
	4	4	4992	12.5
2	1	2	2496	21.2
(RECOMMENDED)	2	4	4992	25
	3	6	7488	25
	4	8	9984	25
4	1	4	4992	42.4
	2	8	9984	50
	3	12	14976	50
	4	16	19968	50
8	1	8	9984	84.8
	2	16	19968	100
	3	24	29952	100
(MAXIMUM)	4	32	39936	100

Note that when the number of disk drives per drive controller increases from one to two, the data transfer rate more than doubles. This is because with one drive per controller, the data transfer rate is limited to the transfer rate of the disk, which for long transfers is 10.6 MByte/s. With two or more disks per controller, however, data files can be interleaved between the disks, thus masking the track-to-track seek time from the transfer process. In this way, the effective transfer rate is the single cylinder transfer rate of the disk drive, which is 12.5 MByte/s for the Ibis 1400.

The effect of improved data transfer rate on MPP system efficiency is shown in Table 2. For purposes of this example, the data base chosen was a Landsat Thematic Mapper image, run through the GAC-developed thematic mapper geometric correction algorithm. The size of the data base is 40 million pixels, each containing seven bytes of information. This data base must be read once, processed, and the results written back to the disks. Total MPP procesing time is projected to be 20 seconds per image. The chart shows the relative overhead imposed by the disk I/O rate on the system's performance, and the time required to load the data for 1 day's (8 hours) worth of processing time. (Note that the time required for the reading and writing of temporary files and ancillary data is ignored.)

TABLE 2: RELATIVE I/O OVERHEAD FOR VARIOUS CONFIGURATIONS

DRIVE TRANSFER RATE (MBYTE/S)		RELATIVE OVERHEAD (%)	I/O TIME REQ'D FOR 8 HRS PROCESSING (HOURS)
1	(PRESENT SYSTEM)	2800	224
10.6	(MIMIMUM SYSTEM)	264	21
25	(RECOMMENDED SYSTEM)	112	9
50		56	5.5
100	(MAXIMUM SYSTEM)	28	2

1.4.4 HARDWARE

The MPP Disk Subsystem, shown in Figure 1, successfully addresses the I/O bottleneck with a highly modularized, expandable system which is capable of storing up to 39,936 MBytes of data, and transferring data into the MPP Staging Memory at a rate of up to 100 MBytes/sec. In the minimum configuration, the subsystem is capable of storing 1248 MBytes, and transferring data to the stager at 10.6 MBytes/sec.

1.4.4.1 Disk Drives

Mass storage for subsystem is provided by commercially available Ibis Model 1400 disk drives. These drives are capable of being controlled by two ports which provides a path for loading and unloading data sets to the drives without loading down either the MPP or the host VAX.

1.4.4.2 Drive Controllers

The disk drives are managed by microprocessor-based drive controllers. In addition to the conventional tasks of interpreting commands and transferring data between system components, these controllers also contain a large (2.7 MByte) data buffer which is used for "masking out" the seek time of the drive and also for "de-skewing" data between the drives. The controllers also have access to a catalog RAM, which contains a copy of the volume information on the disk. By using this catalog, the controllers can more quickly access a given file, without going through the intermediate step of accessing the volume information on the disk.

1.4.4.3 Controller/Stager Interface

The drive controllers communicate to the MPP Staging Memory through the Controller/Stager Interfaces. These interfaces convert the data to the proper logic levels, queue up data between the stager and the controllers, and handle the stager interface protocol.

1.4.4.4 Controller Command Bus

The drive controllers receive command information over the Controller Command Bus (CCB). This bus is really an extension of the Stager Command Bus (SCB) already present in the MPP. Over this bus, the drive controllers can receive commands either from the host VAX or from the MPP I/O Control Unit (IOCU). Receiving commands directly from the IOCU is significant because in this way the MPP can issue requests for data transfers without going through the host computer, thus avoiding the system overhead that these host interrupts would otherwise cause.

1.4.4.5 Controller Data Bus

When not transferring data between the disk drives and the stager, the drive controllers have the capability of transferring data between the disk drives and the host VAX over the Controller Data Bus. The Controller Data Bus is an extension of the VAX DR780 Data Interconnect, with provisions for allowing multiple end devices on the bus.

1.4.5 PHYSICAL CONFIGURATION

The disk subsystem consists of a drive controller cabinet, located next to the MPP chassis, and as many disk drives as required for the desired system configuration. Figure 2 shows one possible layout for the maximum system, including a secondary

VAX-11/780 computer. The disk drive numbers shown in Fig. 2 are follow the same convention as Fig. 1: drive 30 is the first disk drive in row 3. Note that even for the maximum configuration, no disk drive will be located more than 12 feet from the controller chassis, thus easily meeting the 40 foot cable length restriction on the Ibis disk drive control cables.

The drive controller cabinet is the same as was used for the MPP. This will lead to a more aesthetically pleasing appearance for the cabinet, and save on mechanical design costs. The drive controller cabinet is divided up into four quadrants, as is the MPP cabinet. The recommended system will occupy one quadrant, leaving the other quadrants avaliable for future expansion. Each quadrant is capable of holding 26 boards. Total board requirements for some systems is shown in Table 3.

TABLE 3: SUBSYSTEM BOARD REQUIREMENTS

BOARD TYPES	MINIMUM SYSTEM	RECOMMENDED SYSTEM	MAXIMUM SYSTEM
DRIVE CONTROLERS			
CONTROL PROCESSOR	2	8	32
BUFFER MEMORY	2	8	32
VOLUME CATALOG	1	2	8
CONTROLLER/STAGER I/F	1	2	8
SCB-CONT. CMD. BUS I/F	1	2	2
DR780-CONT. DATA BUS I/F	1	2	2
ACCESS RESOLVER	0	1	1
TOTAL	8	25	85

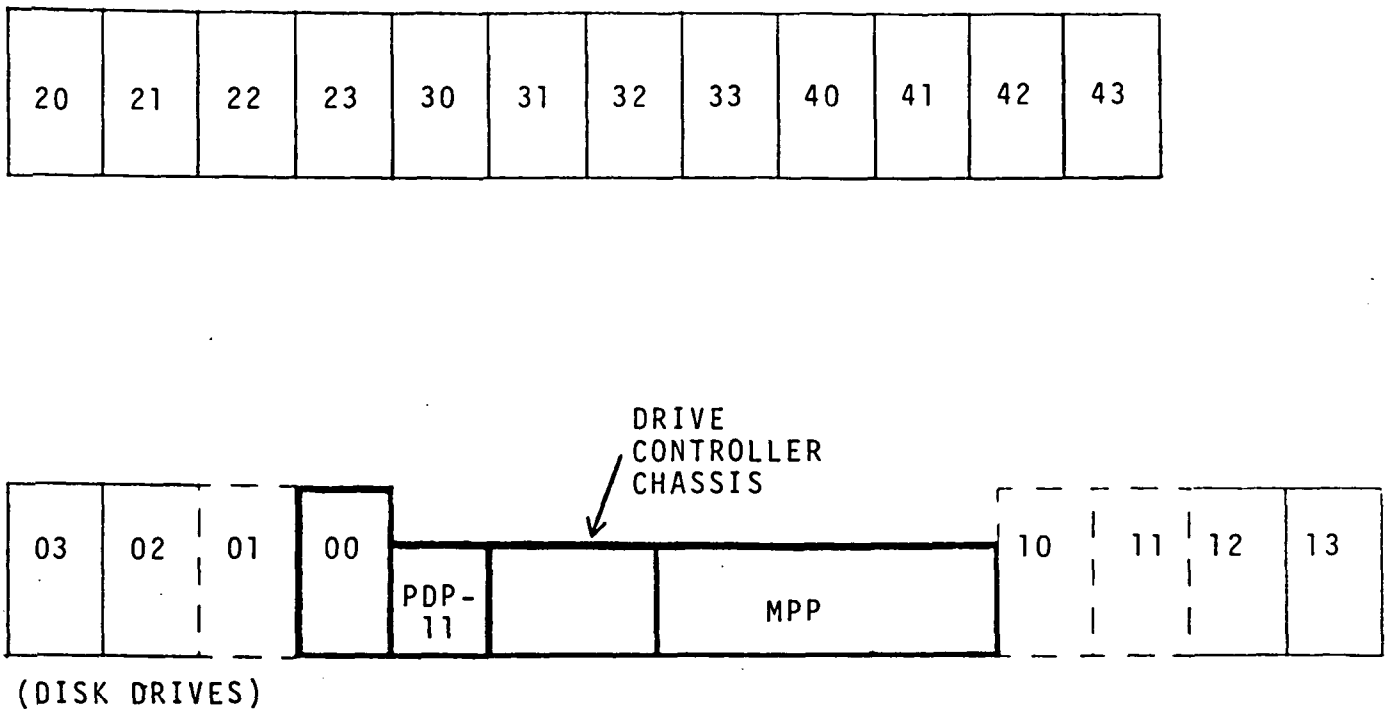


Figure 2. MPP Disk Subsystem Physical Layout

1.5 ALTERNATE SYSTEM

The modular system design described above provides the basis for realizing an MPP disk subsystem in a variety of configurations and associated performance levels. One drawback of the design approach is that realization of the minimum configuration would entail significant non-recurring cost expenditures. In an effort to provide for an MPP disk system at relatively low cost, GAC has investigated an alternate aproach. Basically, the alternate approach attempts to minimize non-recurring costs by capitalizing on disk system components which have recently become available in the commercial marketplace. Non-recurring costs are still required for an MPP/disk system interface and system checkout, but they are much less than for a complete hardware/software system design. The alternate approach allows for a disk subsystem which provides from 1 to 3 GBytes of storage capacity and a data transfer rate of 10 MBytes/s. System storage capacity and transfer rate can be increased.

Realization of the alternate system in either a minimal initial configuration or expanded configuration appears to be straightforward. Confirmation of the viability of the approach, however, will require additional analysis beyond the scope of the present program.

The alternate system is described in Appendix C.

1.6 REFERENCE DOCUMENTS

THEORY OF MPP HARDWARE OPERATION (GER-17143)
MPP STAGING MEMORY (GER-16964)
IBIS MODEL 1400 DISK DRIVE SPECIFICATION
DR-780 USER'S GUIDE (DEC P/N ER-DR780-UG-002)

2. DISK SUBSYSTEM OVERVIEW

The MPP disk subsystem, shown in Figure 1, is built around the MPP host VAX-11/780 computer, which was chosen since it provides an existing interface to the MPP and to the VAX cluster interface, which is scheduled to be installed at NASA in the second half of 1984. Note that the system is expandable in both the horizontal and vertical directions. Expanding horizontally increases storage capacity, but does not increase system transfer rate. Expanding vertically increases both capacity and transfer rate, since the number of bits being transferred to the stager in parallel increases with the number of rows of disk drives. The mimimum system is configured with only drive 03 in place. For added capacity, drives 00 - 02 may be added later. Or, if the user preferred, effective transfer rate and capacity may be increased by adding drives 13 - 73. Total system capacity is 4 rows by 8 columns = 32 drives. This yields a total data capacity of 39936 MBytes of formatted data, and a transfer rate of 100 MByte/sec to the stager.

The disk subsystem consists of four major parts: (1) the disk drives themselves, (2) the drive controllers, (3) the interfaces between the drive controllers and the staging memory, and (4) the buffers between the controller data and command busses and the MPP Stager Command Bus.

The disk drives are Ibis Model 1400 units. These drives have a formatted capacity of 1248 MByte and a transfer rate of 10.6 MByte/s. These drives are commercially available and represent the best combination of capacity and data transfer rate on the market today.

The drive controllers are microprocessor based designs. The controllers consist of two processor boards, two data buffer boards, and one catalog RAM board. The functionality of these boards will be discussed in Section 3.

The Controller - Stager I/F boards are used to buffer the data coming from the drive controllers to the MPP Stager I/O boards. One Controller - Stager I/F board is required per row of disk drives.

Two boards are required to interface the command and data busses of the present system to the drive controlers. One connects the Controller Command Bus (CCB) to the MPP Stager Command Bus (SCB). The other connects the Controller Data Bus to the DR780 Data Interconnect.

A total of seven new designs will be required for implementation of the minimum system: the controllers (2 boards), the data buffers, the catalog RAM, controller-to-stager interfaces, the SCB to CCB interface, and the DR780 to controller data bus interface. Additionally, the MPP Stager I/O boards will be modified to allow connection of the interfaces to the Stager

I/O port. If the recommended system is chosen and the secondary VAX computer is used, an "access resolver" board will need to be designed, in order to convert the DR780 Control Interconnect to the MPP Stage Command Bus. This design will essentially be a re-layout of the SCB resolver now in the MPP. Support hardware necessary for system development and installation includes a motherboard, two extender cards, and a card test adapter.

Figure 1 shows two sets of drive controllers connected to each row of disk drives. The "A" controllers are required for transferring data between the stager and the disk drives. The "B" controllers are required to implement the recommended configuration, and are used for transferring data between the drives and the host (or other) VAX, in a "production" environment. In this way, the loading and unloading of the data sets on the drives can be accomplished without loading down the MPP host VAX.

Note that the subsystem components present in the minimum configuration are highlighted by bold outlines, and the components present in the recommended configuration are set off by dotted lines.

3. DRIVE CONTROLLERS

The drive controller shown in Figure 3 consists of five boards: the two control processor cards, the two buffer memory cards, and a catalog RAM card. The drive controllers are contained in the drive controller cabinet. Interfaces to the Staging Memory Interfaces, the Controller Command Bus, the disk drives, and the Controller Data Interconnect are contained on the two control processor cards.

Each disk drive may be connected to one or two drive controllers. The "A" controller is required and may be used to transmit data between the drives and the stager, or between the drives and the host VAX. Commands to the controllers come from the host VAX, the MPP I/O Control Unit (IOCU), or the PDP-11/34 through the DR11B interface. Contention between various units trying to gain control of the SCB is handled by the SCB resolver in the MPP.

The "B" side drive controller is used primarily for the loading and unloading of data sets between the secondary VAX and the disk drives. Note that since the "B" and "A" controllers will be sharing interleaved access to the same drive, the volume catalog RAM for each row of drives will be shared between the controllers.

3.1 DRIVE CONTROLLER OPERATIONS

* TRANSFER DATA FROM HOST TO DISK
* TRANSFER DATA FROM DISK TO HOST
* TRANSFER DATA FROM DISK TO STAGER ("A" SIDE ONLY)
* TRANSFER DATA FROM STAGER TO DISK ("A" SIDE ONLY)
* TRANSFER DATA FROM AUX. PORT TO DISK ("B" SIDE ONLY)
* TRANSFER DATA FROM DISK TO AUX. PORT ("B" SIDE ONLY)
* TRANSFER DATA FROM HOST TO STAGER ("A" SIDE ONLY)
* TRANSFER DATA FROM STAGER TO HOST ("A" SIDE ONLY)
* RECEIVE COMMAND STRING FROM HOST
* PERFORM DIAGNOSTIC AND TEST COMMANDS

3.1.1 WRITE OPERATION

When writing data to the disk(s), data will be written to the data buffer while the seek command is in execution. While the disk is seeking, the interface is permitted to fill up both halves of the data buffer. Once the disk reports "on cylinder", it is commanded to accept data, starting at the next sector. In this way, the rotational latency of the drive is reduced or eliminated. As soon as the first cylinder is depleted, the drive will seek to the next cylinder, and the interface is permitted to write again to the first buffer. This process continues until the

write operation is completed.

3.1.2 READ OPERATION

The read operation works similarly to the write operation, except that the interface is not permitted to take data from the data buffer until the first cylinder's worth of data has been loaded.

3.2 MAJOR CONTROLLER COMPONENTS

3.2.1 DATA BUFFERS

The data buffers are used in this system to eliminate the rotational latency inherent in any rotating media. When more than one row of drives is present, the data buffers also serve to deskew the data between the disk drives.

The data buffers are used to transfer data between the drive controller and the controller-to-stager interfaces. When writing data to the disk, the transfer to the buffers can begin immediately, without waiting for the drive seek to complete. When reading data from the disk, the data buffer can begin to be filled as soon as the drive is on cylinder, without waiting for sector 0 to be read. As soon as one cylinder has been read, that half of the data buffer is available for transfer to the controller-to-stager interface, while the drive controller is transferring the next cylinder of data into the other data buffer half.

The size of this buffer is determined by the number of bytes per cylinder for the drive in use. The buffer must accomodate two cylinders worth of data. For the IBIS drive, this buffer must be 2.7 MByte. In operation, the controller will begin to load the buffer immediately when the seek command is complete. When the entire cylinder is read, the drive seeks to the next cylinder and begins to fill the other half of the buffer. Data transfers to the stager may begin as soon as all drives have completed the read of the first cylinder.

This memory will be built with 256k RAMs, and will use error correction logic to correct single bit errors, and detect double bit errors.

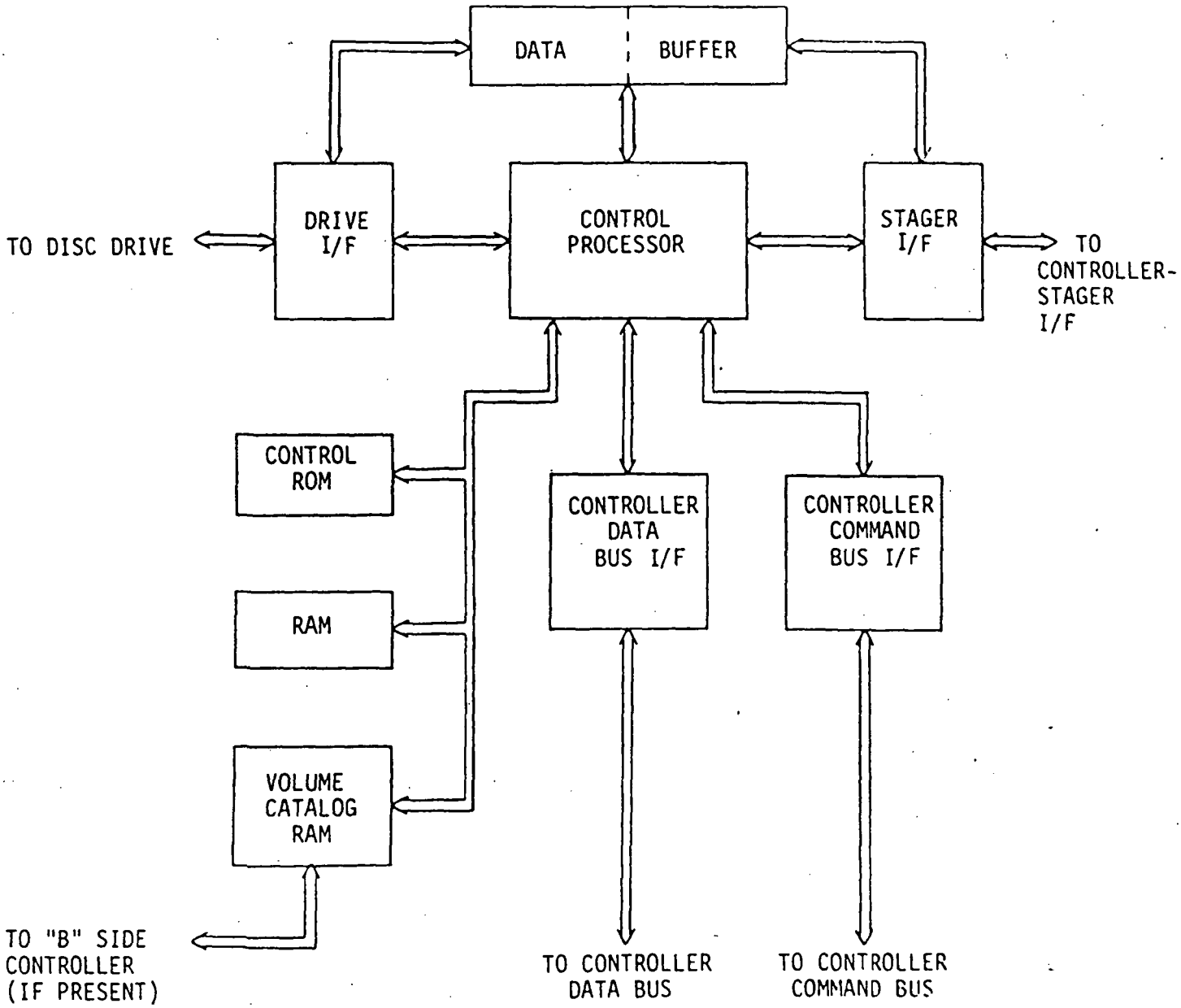


FIGURE 3: DRIVE CONTROLLER BLOCK DIAGRAM

3.2.2 CATALOG RAM

The drive controllers will maintain in their own RAM, copies of the volume catalog stored on the drive(s). This RAM will be shared between the "A" and "B" controllers for each row of disk drives, so that file additions or deletions by either controller will be immediately available to the other. By using this catalog RAM, the subsystem will be able to avoid disk reads when locating files. The controllers will thus have the capability of accessing files by file name, rather than by track and sector number. In order to further minimize accesses to the volume catalog, this RAM need not be updated when the controller is dealing with temporary files. This memory will also be built with 256k RAMs, and will use error correction logic.

3.3 CONTROLLER INTERFACES

The drive controllers must interface to the following devices:

* Staging Memory Interfaces
* Disk drives
* Controller Data Bus
* Controller Command Bus

3.3.1 STAGING MEMORY INTERFACES

Each drive controller communicates with its associated staging memory interface over the following lines:

CONTDATA<0-15,P>-1	BIDIRECTIONAL, 16 BIT DATA BUS, ODD PARITY
IFQFULL-1	BIDIRECTIONAL SIGNAL, SENT BY THE RECEIVER, TO INDICATE TO THE DRIVER THAT IT SHOULD TEMPORARILY STOP SENDING DATA
IFQWRTCLK-1	BIDIRECTIONAL WRITE CLOCK DRIVEN BY SENDING DEVICE.
READCMD-1	DRIVEN BY CONTROLLER, COMMANDS I/F TO BEGIN READING DATA FROM THE STAGER
WRITECMD-1	DRIVEN BY CONTROLLER, COMMANDS I/F TO BEGIN WRITING DATA TO THE STAGER

3.3.2 IBIS DISK DRIVE INTERFACE SIGNALS

(A more complete description of this interface is given in Appendix A.)

TABLE 4: IBIS DISK DRIVE INTERFACE SIGNALS

SIGNAL	DRIVER	MEANING
BUS<0-15,P>-1	BOTH	SIXTEEN BIT DATA BUS, ODD PARITY
CODE<0-3,P>-1	CONT.	THREE BIT CMD/STATUS CODE, ODD PARITY
FUNCTION READY-1	CONT.	INDICATES THAT CODE<0-3> IS VALID
READY-1	DRIVE	INDICATES THAT ENABLED DRIVE IS READY
RDCLK-1	DRIVE	100 NS DATA CLOCK
ERROR-1	DRIVE	INDICATES ERROR OR DRIVE FAULT STATUS
WRCLK-1	CONT.	100 NS DATA CLOCK
SELECTED-1	DRIVE	INDICATES THAT DRIVE IS SELECTED
BUSY-1	DRIVE	INDICATES THAT SELECTED DRIVE IS BUSY
BUSSAFE-0	CONT.	HIGH INDICATES OPEN CABLE TO DRIVE
DATAREQ-1	BOTH	INDICATES THAT RECEIVING DEVICE IS READY FOR MORE DATA
DIRIN-1	CONT.	LOW INDICATES THAT CONT. IS DRIVING THE DATA BUS
RESET-0	CONT.	LOW RESETS ALL DRIVES ON THE BUS
STATUSP-1	DRIVE	ODD PARITY OF READY, ERROR, SELECTED, AND BUSY
DEVENB<0,1>-1	CONT.	USED TO ENABLE 1 OF 4 DRIVES ONTO BUS
DATARDY-1	BOTH	INDICATES THAT DATA ON BUS IS VALID

3.3.3 CONTROLLER COMMAND BUS INTERFACE SIGNALS

The Controller Command Bus (CCB) is the "decoupled" SCB which connects the drive controllers to the Stager Command Bus. Timing for the CCB is similar to that for the SCB. The CCB connects 1, 2, 4, or 8 drive controllers to the SCB. The purpose of the CCB is to transmit commands from the MPP or host to the controller(s), and to return status from the controller(s) to the host.

TABLE 5: CONTROLLER COMMAND BUS INTERFACE SIGNALS

SIGNAL	DRIVER	MEANING
CCBMSTRSYNC-1	I/F	SYNC CLOCK, DRIVEN BY BUS MASTER
CCBSLVSYNC-1	CONT.	SLAVE CLOCK, DRIVEN BY ENABLED CONT.
CCBFUNCT<0,1>-1	I/F	INDICATES OPERATION IN PROCESS
CCBSENSE-1	CONT.	INDICATES SENSE OF FLAG BIT, BAD PARITY, OR INVALID ADDRESS
CCBDATA<0-7,P>-1	BOTH	EIGHT BIT BUS FOR COMMANDS AND STATUS
CONTERRINT<0-7>-0	CONT.	EACH CONTROLLER MAY DRIVE ONE OF THESE LINES TO INDICATE THAT AN ERROR HAS BEEN DETECTED.
CONTSTATINT<0-7>-0	CONT.	EACH CONTROLLER MAY DRIVE ONE OF THESE LINES TO INDICATE THAT A STATUS INTERRUPT HAS OCCURRED.

3.3.4 CONTROLLER DATA BUS SIGNALS

The drive controllers will connect to the VAX DR-780 Data Interconnect (DI) bus through an interface card which will be located in the drive controller cabinet next to the MPP. Due to the size of this design, this interface will share a card with the CCB - SCB interface, above. Note that since the controllers work on a sixteen bit data bus, bits 16 - 31 on the DR780 will be ignored. (The DR-780 can be configured for this.) All signals on the CONTDI bus will be differentially driven, RS-422 levels.

TABLE 6: CONTROLLER DATA BUS SIGNALS

SIGNAL	DRIVEN BY	MEANING
CONTDICKAB-1	DR780	DATA CLOCK, DRIVEN BY DR780
CONTDICKBA-1	CONT.	DATA CLK, DRIVEN BY CURRENT SLAVE
CONTDISEND<2-0>-1	CONT.	INDICATES ENCODED SENSE OF SEND, DATA, AND VALID DATA POSITIONS
CONTDID<0-15,P>-1	BOTH	BIDIRECTIONAL DATA BUS
CONTDIDIREC-1	BOTH	INDICATES BUS DIRECTION
CONTDIRRDY-1	BOTH	INDICATES THAT RECEIVER IS READY

3.3.5 EXTERNAL PORT INTERFACE

In addition to the data ports described above, the "B" side

controllers will be capable of interfacing to an additional high speed (20 MByte/sec.) data port. It is expected that this port will be used to connect the controllers to high-density tapes or other primary media. The protocol and timing for this port is the same as for the Staging Memory Interface, above. In this way, the external port can be implemented without impacting system design, and the "B" side controllers can be completely identical to the "A" side controllers.

4. CONTROLLER - STAGER INTERFACE

These cards (one per row of drive controllers) are used to buffer data between the controllers and the MPP Staging Memory I/O cards. The staging memory interfaces contain logic to handshake both with the drive controllers and the Stager I/O cards, and a limited buffer for the data being transferred. These cards also contain logic to communicate between each other, in order to pass information about whether or not data is ready for transfer, as well as data parity.

The disk subsystem is designed to interface directly with the staging memories of the MPP. The subsystem can be configured with 1, 2, 4 or 8 rows of drives which means that the data channel into the stager will be 16, 32, 64 or 128 bits wide. If the data channel is less than 128 bits wide, the data into or out of the stager will be reconfigured using the flip network within the stager.

Since all bits are not present on all stager I/O boards, the I/O channel from each drive controller must be spread among multiple stager boards. The algorithm for determining this data spread is as follows: Each controller is assigned a number from 0 - 7, which is represented as: C(4)C(2)C(1). The data bits from each controller is assigned a number from 0 - 15, which is represented as: D(8)D(4)D(2)D(1). Each bit may be assigned to a particular bit from 0 - 15 on a stager I/O board from 0 - 7, as follows:

Stager I/O Board Number = D(8)D(4)C(1)
Data Bit Number = C(4)C(2)D(2)D(1)

By applying this formula, the data bits from each drive controller can be allocated to the stager I/O boards, according to Table 7:

TABLE 7: DRIVE CONTROLLER BIT FANOUT

CONTROLLER	BITS	GO TO STAGER I/O BOARD	BITS
0	0-3	0	0-3
	4-7	2	0-3
	8-11	4	0-3
	12-15	6	0-3
1	0-3	1	0-3
	4-7	3	0-3
	8-11	5	0-3
	12-15	7	0-3
2	0-3	0	8-11
	4-7	2	8-11
	8-11	4	8-11
	12-15	6	8-11
3	0-3	1	8-11
	4-7	3	8-11
	8-11	5	8-11
	12-15	7	8-11
4	0-3	0	4-7
	4-7	2	4-7
	8-11	4	4-7
	12-15	6	4-7
5	0-3	1	4-7
	4-7	3	4-7
	8-11	5	4-7
	12-15	7	4-7
6	0-3	0	12-15
	4-7	2	12-15
	8-11	4	12-15
	12-15	6	12-15
7	0-3	1	12-15
	4-7	3	12-15
	8-11	5	12-15
	12-15	7	12-15

Placing data directly into the stager will require some modification to the stager I/O boards. Specifically, the external output port drivers of the stager must be tied to the stager output queue, and the stager input queue must be tied to the stager input data selector. Also, some minor logic changes must be made in the queue control logic of the stager.

5. DR780-CONTROLLER DATA BUS INTERFACE

This interface consists of a single card which allows multiple drive controllers to have (non-simultaneous) access to the Data Interconnect channel of the DR780 interface. Signals on the DI side of the interface may be single- or diffenentially-driven. All signals on the CONTDI side of the interface are differentially driven, RS-422 level signals.

Note that both "A" and "B" side drive controllers may be tied to the same controller data bus.

5.1 DR780-CONTROLLER DATA BUS SIGNAL LEVELS

TABLE 8: DR780-CONTROLLER DATA BUS SIGNAL LEVELS

DR780 SIGNAL	TYPE	CONTDI SIGNAL
DATA<31:0,P>-0	SINGLE	CONTDID<0-15,P>-1
CKAB-1	DIFF.	CONTDICKAB-1
CKBA-1	DIFF.	CONTDICKBA-1
RECEIVE	DIFF.	CONTDIREC-1
RRDY	DIFF.	CONTDIRRDY-1
SEND<2:0>-0	SINGLE	CONTDISEND<2-0>-1

In this application, CKAB is driven by the host DR780, and CKBA is the return clock from the enabled controller.

5.2 DATA TRANSFER OVER DATA INTERCONNECT

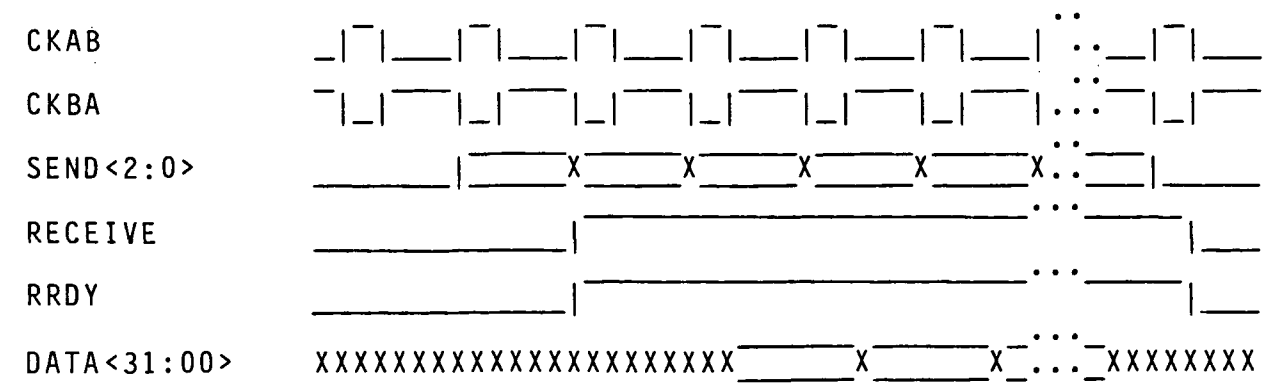


FIGURE 4: DATA TRANSFER OVER DR780 DATA INTERCONNECT

6. CONTROLLER COMMAND BUS TO STAGER COMMAND BUS INTERFACE

6.1 INTRODUCTION

This card is used to place the drive controllers into the MPP Stager Command Bus enviornment, allowing the drive controllers to receive commands and transmit status to the MPP I/O Control Unit and the host VAX.

The drive controllers receive commands and transmit status via the Controller Command Bus (CCB), which is a logical extension of the MPP Stager Command Bus (SCB). The SCB is implemented in the MPP to allow multiple end devices to be connected to the DR780 interface, and to transmit stager commands within the MPP.

The SCB allows up to sixteen stagers, each with up to eight devices. The drive controllers will contain switches allowing them to be configured as any device in any stager. Note that both the "A" and "B" drive controllers may co-exist on the same bus, as long as the switches are set so as to avoid address conflict. Additional switches will also be provided for a "common" address.

The common address is used to transmit one command to all drive controllers simultaneously, as when writing data to the stager. The individual addresses are used for commands to individual drive controllers, such as transferring data from a disk to the host VAX-11/780.

Signals present on the SCB are as follows:

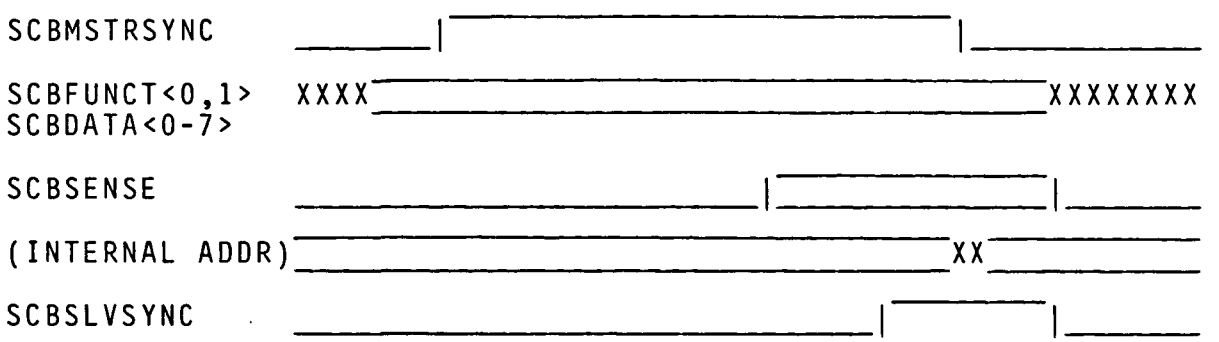
TABLE 9: SCB SIGNAL DEFINITION

SIGNAL NAME	MEANING	
SCBDATA<0-7,P>-1	SCB DATA, ODD PARITY	
FUNCT<0,1>-1	DEFINES SCB FUNCTION:	00 - DEVICE ADDRESS
		01 - READ
		10 - WRITE
		11 - FLAG COMMAND
SCBMSTRSYNC-1	INDICATES THAT DATA ON THE BUS IS VALID	
SCBSLVSYNC-1	ACKNOWLEDGES SCBMSTRSYNC AT FUNCTION COMPL.	
SCBSENSE-1	FLAG BIT, BAD PARITY, OR INVALID ADDRESS	

The "address" transaction is used to enable slave devices onto the bus. Address transactions may be followed by : 'n' commands, two writes followed by reads or writes, or 'n' commands followed by two writes, followed by reads or writes.

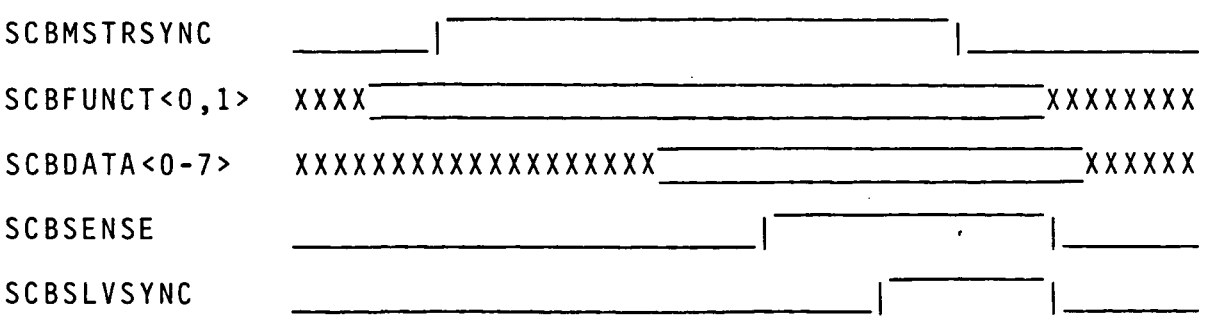
Note that stager 0, devices 3 and 7 are not presently used, so these addresses are also available for use by the controller(s).

6.2 SCB TIMING



SETUP TIME, SCBFUNCT AND DATA PRIOR TO SCBMSTRSYNC	0 ns min
RESPONSE TIME, SLVSYNC HI TO INT. ADDR. INCREMENT	60 ns min
RESPONSE TIME, SLVSYNC HI TO MSTRSYNC LOW	0 ns min
HOLD TIME, MSTRSYNC LOW TO FUNCT, DATA INVALID	0 ns min
HOLD TIME, MSTRSYNC LOW TO MSTRSYNC LOW	0 ns min

FIGURE 5: WRITE, COMMAND AND ADDRESS TRANSACTIONS OVER SCB



SETUP TIME, SCBFUNCT PRIOR TO SCBMSTRSYNC	0 ns min
SETUP TIME, SENSE HI PRIOR TO SLVSYNC HI	60 ns min
RESPONSE TIME, SLVSYNC HI TO MSTRSYNC LOW	0 ns min
HOLD TIME, MSTRSYNC LOW TO FUNCT, DATA INVALID	0 ns min

FIGURE 6: READ TRANSACTIONS OVER SCB

7. SYSTEM REQUIREMENTS/IMPLICATIONS

7.1 FILE LENGTH, LOCATION

All data files on the drive(s) will be restricted to an integral number of cylinders. If a file fills up only part of a cylinder, the remainder of the cylinder is not used. Also, the drive controller will attempt to keep all files located contiguously on the drive(s). In this way, seek times will be minimized.

7.2 SOFTWARE REQUIREMENTS

Since the Ibis disk drives do not emulate DEC drives, new driver software must be written for the host VAX computer. Also, new file management software must be written for the VAX to handle the spreading of the data between multiple disk drives.

The Staging Memory Manager (SMM) in the MPP must be modified, in order to handle the new data spreading pattern.

In addition, diagnostic and test programs will need to be written for both the VAX and the drive controller. Finally, control firmware for the drive controller processor and its I/O state machines must be written.

7.3 DEGRADED MODE OPERATION

In the event that a disk drive should fail, the subsystem is able to continue operating merely by not using that disk. If more than one drive should fail in different rows, operations can still continue by reconfiguring the disks so both defective drives are in the same column. This can be done via a "DIP" switch which is located on the disk drive's input/output card.

7.4 HOST VAX LIMITATIONS

Figure 1 shows the possibility of connecting two DR780 interfaces to the MPP host VAX-11/780. While this can be done, experience has shown that doing so seriously degrades the performance of the VAX. In a "batch load" environment where the data sets would be loaded into the disk farm during off hours, processed during the day, and offloaded during off hours, the data may be loaded without seriously degrading system performance. However, in a "production" environment, where the data sets will be loaded while the MPP is busy, using the host VAX to load data sets would degrade system performance. In this case, it is recommended that the "B" ports of the Ibis drives be connected, through their controllers, to a second VAX computer, which would be used exclusively for loading and unloading data

sets.

7.5 HOST INTERRUPTS

The subsystem will have the capability of presenting error and status interrupts to the host, if enabled. Error interrupts will be assigned to interupt level 6, and status interrupts will be level 7. Note that the host will need to poll the controllers in its interrupt handler to determine which controller presented the interrupt, and to check for possible mutiple interrupts.

The interrupt signals are presented to the MPP Power Sequencer board, and are ECL level, positive true. The host interrupt vector for the error interrupt is octal 660, while for the status interrupt it is octal 670.

8. ALTERNATIVE TECHNOLOGIES EXPLORED (SUMMARY)

A number of alternative technologies, beside disks, were explored. Basically, the other technologies fell short of requirements due to cost, lack of availability, insufficient speed, or poor reliability. A chart summarizing the alternatives appears in Appendix A. A discussion of the alternatives appears below.

8.1 SOLID-STATE DISKS

Solid-state disks, also known as disk emulators and RAM caches, possess some unique advantages over conventional disks. Primarily, these advantages are the elimination of seek time and rotational latency delays from the system. Additionally, the data path and cache size can be easily expanded to allow virtually unlimited transfer rates and capacities. The primary disadvantages of the RAM approach to data storage are cost, power consumption, volatility of the data, and reliability. Most solid state disks in the marketplace, with the exception of the STC 4305, are aimed at the mini- and microcomputer marketplaces. The STC 8890 (CyberCache) takes a hybrid approach by placing frequently used files in RAM and otherwise functioning as a normal disk drive.

The hardware cost of a 1GByte RAM array has been estimated to be $1.2M. This estimate assumes the use of 128k RAMs costing $15 each, using 7 ECC bits over 32-bit words. Such a system would dissipate 2 kW, and would have an MTBF of 83 hrs between (recoverable) data errors, and 250 hrs between (recoverable) memory device failures. The reliability figures are based on an assumed reliability rate of 300 soft errors per billion device hours, and 100 hard errors per billion device hours.

8.2 BUBBLE MEMORIES

Bubble memories offer some advantages over dynamic RAMs, namely nonvolatility of data and increased bit densities. However, bubble memories are not as widely available as RAMs, are more expensive, and have totally unacceptable access times. Typical access times for bubble memories are 40 milliseconds.

8.3 OPTICAL DISKS

Optical (laser) disks offer the significant advantages of very high bit densities and excellent long term storage characteristics. Transfer rates for these systems are extremely slow, however. (500kByte/sec for the Shugart Optimem 1000, for instance.) Additionally, most of these systems are just beginning to start up into production quantities.

8.4 SERIAL DISKS

Most common in the marketplace are the serial transfer disk drives. Since competition is high, prices are low, and drive capacities are steadily increasing, and there is much standardization. Most of the drives are 14 inch Winchesters, using the Storage Module Drive (SMD) interface. The best performer in this arena appears to be the Fujitsu Eagle (M2351A), which is a 10.5-inch Winchester offering 474 MBytes of unformatted storage, and a burst transfer rate of 1.92 MByte/sec.

The biggest disadvantage of these drives, aside from their low transfer rates, is that the interface to the drive requires the use of complex linear circuitry to decode the data. More than one vendor has cautioned against underestimating the complexity of the task of designing this interface.

8.5 PARALLEL DISKS

The drives with the highest performance characteristics are the parallel transfer disk drives. These drives are capable of transferring eight (or more) bits to the host simultaneously. The main contenders in this field are the AMPEX 9309 and the IBIS 1400.

The Ibis 1400 disk drive contains nine non-removable platters of thin film media. The drive contains 1.2 GBytes of formatted data and transfers data at an average rate of 12.5 MByte/s for transfers of one cylinder (1.375 MByte) or less, and 10.6 MByte/s for transfers two cylinders or longer. The drive contains an internal control card which presents the data to the external interface in a completely synchronous manner.

The Ampex 9309 drive has an unformatted capacity of 312 MBytes and transfers data at an average rate of 7.8 MBytes/sec. AMPEX also sells the controller for the drive, which is called the DCP-909. Each controller is capable of supporting up to four drives. The only currently available host interface for the DCP-909, however, is to a DEC UNIBUS. Also, the Ampex drive is not dual ported, so the secondary VAX loading feature of the recommended system could not be used if the Ampex drive were chosen.

After examining all the alternatives, the IBIS 1400 drive was selected as the drive of choice, as it presented the best mix of capacity, data transfer rate, cost and reliability, and was also dual ported. The dual port feature of the drive was needed to provide a path for data from a secondary VAX (or other) computer. The Ibis drive has another advantage in that it's interface is completely synchronous. This will help reduce the cost and improve the reliability of the interface hardware.

9. CONCLUSIONS

9.1 INTRODUCTION

The disk subsystem described in this document will provide NASA with a flexible mass memory posessing both large capacity and high I/O data rates. The system is easily expandable in terms of capacity, data rate, and data sources/sinks. The system requires no new technology, only utilization of commercially available disk drives and application of sound hardware and software design techniques. In addition, the subsystem can be implemented with minimum modification of existing MPP hardware and software.

9.2 SYSTEM PERFORMANCE VERSUS CONTRACT SPECIFICATION

The system design meets performance specifications, as shown by Table 10.

TABLE 10: SUBSYSTEM PERFORMANCE VS. CONTRACT SPECIFICATION

ITEM	SPEC	DESIGN
CAPACITY (INITIAL)	1 GBYTE	1.2 GBYTE
CAPACITY (FINAL)	5 GBYTE	40 GBYTE
TRANSFER RATE TO STAGER (INITIAL)	10 MBYTE/S	10.6 MBYTE/S
TRANSFER RATE TO STAGER (FINAL)	40 MBYTE/S	100 MBYTE/S
TRANSFER RATE TO VAX	6.6 MBYTE/S	6.6 MBYTE/S

9.3 SYSTEM FEATURES SUMMARY

The disk subsystem, as designed, features a highly modularized, expandable disk "farm" built around the Ibis Model 1400 disk drive. The subsystem contains intelligent drive controllers which are capable of accessing files by file name, thus reducing the processing load on the host VAX-11/780 computer. The controllers also contain a large data buffer, which is capable of masking out the rotational latency of the rotating media. The disk subsystem makes efficient use of parallelism by accessing up to eight disks simultaneously to provides throughput rates far beyond those of conventional disk systems.

9.4 DESIGN COMPLETION REQUIREMENTS

9.4.1 SYSTEM DESIGN

The disk subsystem is presently designed to a top-level block diagram. Completion of the design would require a more detailed design of the interface protocol between the blocks, as well as study of the hardware partitioning of the design. In addition, some software systems analysis must be performed. At that point a detailed block diagram for each major system component can be designed, and the performance of the entire subsystem can be analyzed and verified.

9.4.2 HARDWARE DESIGN

Once system design is complete, detailed hardware design can begin for the drive controllers, the controller to stager interfaces, the DR780 to Controller Data Interconnect interface card, and the SCB to CCB interface card. Of these, the design for the drive controllers is believed to be the most complex, since it involves microprocessor design, large RAM arrays, and several state machines to interface with other blocks within the system.

9.4.3 SOFTWARE DESIGN

Software design requirements would be for the drive controller firmware, a new file management module, new VAX drivers, modification to the current Staging Memory Manager, and various test and diagnostic routines.

9.4.4 PROCUREMENT

Parts requirements for the recommended system include four Ibis disk drives, approximately 15 MBytes worth of dynamic RAM, and a VAX-11/780 computer. Delivery requirements for the system components would be spread out over the life of the design cycle. For instance, since detailed software design can begin without the availability of the Ibis drive or its controller, the VAX computer would be delivered early in the design phase. On the other hand, since only one disk drive is required for debug virtually throughout the design cycle, the remaining drives need not be delivered until relatively late in the cycle. By intelligent management of equipment deliveries, costs can be more accurately controlled.

9.5 INSTALLATION REQUIREMENTS

The system is configured such that the drive controller cabinet must be located adjacent to the MPP chassis. The disk drives can be located up to 40 feet (cable length) from the

controller cabinet. In addition, the drives and controller cabinet each require 30 inches front and rear clearance for access to the units.

9.5.1 IBIS MODEL 1400 DISK DRIVE

POWER	SIZE
208 VAC +/- 10%	24 IN. WIDE
3 PHASE DELTA	44 IN. DEEP
5 WIRE CABLE	54 IN. HIGH
20 AMP SERVICE	844 LB.
60 HZ.	

9.5.2 CONTROLLER AND INTERFACE CHASSIS

POWER	SIZE
120 VAC +/- 10%	24 IN. WIDE
60 HZ.	35 IN. DEEP
10 AMP SERVICE (EST.)	50 IN. HIGH

9.6 SCHEDULE AND COST

The schedule and budgetary cost data shown below represents an estimate as to the time and resources required to complete the task. The budgetary estimate is intended to represent the relative magnitude of the effort; it is not a price quote for this project.

The schedule and cost data shown below assume an 18 month implementation schedule. The labor charges for design engineers is included in the cost for their part of the design.

	ITEM	COST
1.	MANAGEMENT PROJECT ENGINEER LIASON PERSONNEL ADMINISTRATIVE ENGINEERING	$330K

SUPPORT ENGINEERING

2. SYSTEM DESIGN $100K
   BASIC CONCEPT
   INTERFACES
   CONTROL
   PARTITIONING
   SOFTWARE SYSTEM
   PERFORMANCE VERIFICATION ANALYSIS

3. ENGINEERING DESIGN $160K
   PRINTED CIRCUIT BOARDS
      DRIVE CONTROLLER PROCESSOR #1
      DRIVE CONTROLLER PROCESSOR #2
      DRIVE CONTROLLER DATA BUFFER MEMORY
      DRIVE CONTROLLER VOLUME CATALOG MEMORY
      CONTROLLER TO STAGER INTERFACE
      DR780 TO CONTROLLER INTERFACE
      SCB TO CCB INTERFACE
      ACCESS RESOLVER
      STAGER I/O REDESIGN
      MOTHER BOARD SUPPORT
      CARD EXTENDER
      CARD TESTER ADAPTER BOARD
   CABLES
      POWER SYSTEM
      CONTROL/TEST PANELS
      DOCUMENTATION

4. SOFTWARE DESIGN $400K
   SYSTEM DESIGN
   FILE MANAGEMENT MODULE
   STAGING MEMORY MANAGER MODIFICATION
   MCL MODIFICATION/ENHANCEMENT
   VAX DRIVERS
   MICROPROCESSOR CONTROL FIRMWARE
   STATE MACHINE LOGIC CONTROL
   TEST PLANS
   TEST ROUTINES
   DIAGNOSTICS
   VAX/MPP SUPPORT
   DOCUMENTATION

5. ELECTRICAL PRODUCT DESIGN $200K
   PRINTED CIRCUIT BOARDS (SCHEMATIC AND LAYOUT)
      DRIVE CONTROLLER PROCESSOR #1
      DRIVE CONTROLLER PROCESSOR #2
      DRIVE CONTROLLER DATA BUFFER MEMORY
      DRIVE CONTROLLER VOLUME CATALOG MEMORY

CONTROLLER TO STAGER INTERFACE
DR780 TO CONTROLLER INTERFACE
SCB TO CCB INTERFACE
ACCESS RESOLVER
STAGER I/O REDESIGN
MOTHERBOARD
CARD EXTENDER
CARD TESTER ADAPTER
CABLES
CARD RETAINERS

6. MECHANICAL PRODUCT DESIGN $90K
NEW CABINET
CHASSIS
CARD FRAME ASSEMBLY
POWER SUPPLY
POWER DISTRIBUTION
COOLING

7. OTHER PRODUCT DESIGN $55K
ARTWORK GENERATION
WIRE LIST
LIASON
CHECKOUT SUPPORT
SUPERVISION

8. MANUFACTURING $160K
PCB FABRICATION
DRIVE CONTROLLER #1 (4 EA.)
DRIVE CONTROLLER #2 (4 EA.)
DRIVE CONTROLLER DATA BUFFER MEMORY (4 EA.)
DRIVE CONTROLLER VOLUME CATALOG MEMORY (2 EA.)
CONTROLLER TO STAGER INTERFACE (2 EA.)
DR780 TO CONTROLLER INTERFACE (2 EA.)
SCB TO CCB INTERFACE (2 EA.)
ACCESS RESOLVER (1 EA.)
STAGER I/O BOARDS (9 EA.)
MOTHERBOARD (1 EA.)
CARD EXTENDER (2 EA.)
CARD TESTER ADAPTER (1 EA.)

PCB ASSEMBLY
DRIVE CONTROLLER PROCESSOR #1
DRIVE CONTROLLER PROCESSOR #2
DRIVE CONTROLLER DATA BUFFER MEMORY
DRIVE CONTROLLER VOLUME CATALOG MEMORY
CONTROLLER TO STAGER INTERFACE
DR780 TO CONTROLLER INTERFACE
SCB TO CCB INTERFACE

ACCESS RESOLVER
STAGER I/O BOARDS
MOTHER BOARD
CARD EXTENDER
CARD TESTER ADAPTER

OTHER ASSEMBLY
CABLES
WIRE WRAP
CABINET

OTHER MANUFACTURING
TOOL FABRICATION
PARTS MANUFACTURING
PLANNING
INSPECTION

9. QUALITY ASSURANCE $15K
ENGINEERING
VENDOR SURVEY
INCOMING FUNCTIONAL TEST
PRE-SHIP INSPECTION
DOCUMENTATION

10. CHECKOUT $320K
IN HOUSE
SYSTEM SUPPORT
HARDWARE
TEST PROCEDURES
PCBs
UNIT TEST
SYSTEM TEST
TECHNICIAN SUPPORT
SOFTWARE
TEST PROCEDURES
MODULE TEST
SOFTWARE SYSTEMS SUPPORT
HARDWARE SUPPORT
DOCUMENTATION
ON SITE
HARDWARE
SOFTWARE
TECHNICIAN SUPPORT
SUPERVISION
MATERIAL
PER DIEM
TRANSPORTATION

11. PACK AND SHIP

	LABOR	
	MATERIAL	
12.	OTHER	$810K
	PURCHASED MATERIAL	
	VAX-11/780 SYSTEM WITH DR780	
	IBIS MODEL 1400 DISK DRIVES (4 EA.)	
	PIECE PARTS (LESS MEMORIES)	
	MEMORIES	
	RAW MATERIAL	
	CABINET & ASSOCIATED HARDWARE	
	SUPPORT	
	CAPITAL	
	IBM PC FOR SCHEMATIC GENERATION	
	DESIGN REVIEWS	
	DOCUMENTATION	
TOTAL	LABOR	$1830K
	MATERIAL	$810K
TOTAL FACTORY COST		$2640K
SELL PRICE		$3370K

9.7 TECHNICAL AND COST RISKS

The overall system entails a moderate risk. The primary risks lie in the Ibis disk drives, the drive controllers, and the system software design. The rework of the Stager I/O boards, the Stager-to-Controller Interface, the SCB interface logic, and the Controller data bus-to-DR780 interfaces are all well understood, and entail low technical, cost, and schedule risks.

The drive controllers entail risk primarily due to the complexity of the task. The controllers must connect to the Ibis interface, the controller-to-stager interface, the Controller Command Bus, and the Controller Data Bus. The controllers must contain approximately 5.4 MByte of buffer RAM, and will be microprocessor controlled. It is estimated that the control software for the controllers will run to about 16 KBytes.

The Ibis drives entail risk due to the start-up nature of Ibis Systems, Inc. Ibis is presently a one product company, and although they have been extremely successful in raising venture capital in the past, there is no guarantee for the future.

In their favor, Ibis does have a longstanding relationship with Cray Research. (They share some of the same venture capitalists.) Goodyear Aerospace personnel have visited the Ibis plant. Their facilities and staffing appear to be more than adequate for the task. Also, two Ibis customers (Cray and E-Systems) have been contacted. Both are satisfied with the performance and the reliability of the Ibis drive, and plan to continue to use them in the future.

9.8 CONCLUSIONS AND RECOMMENDATIONS

The MPP Disk Subsystem, as designed, represents a very effective, highly expandable pathway for removing the I/O bottleneck from the current MPP system. This subsystem can be implemented without an extensive redesign of either the MPP or any of its support hardware of software.

The subsystem configuration recommended (four disk drives, four controllers, and a supplementary VAX-11/780) represents the best compromise between implementation cost, performance, and expansion cost. The purchase of the additional VAX means that hardware and software can be checked out at the factory, while the additional disk drives and controllers represent a substantial improvement in performance at relatively low incremental cost over the minimum configuraion.

APPENDIX A. DISK DRIVE COMPARISON SUMMARY

SERIAL TRANSFER DRIVES

MODEL	XFER RATE (MBYTE/SEC)	CAPACITY (MBYTE)	SEEK TIME	PRICE	I'FACE
DEC RA81	2.2	627	6 MS (1 TRK)		PROPRIETARY
CDC HCD9797	4.8	600	50 MS (AVG)	$40-50k	CUSTOM
CDC 9715-500	1.8	516	5 MS (1TRK)		SMD, ISI
CDC XMD (HYDRA)	1.8	825	5 MS (1 TRK)	$12k	SMD, ISI
PRIAM 15450	1.2	158	9.6 MS (1 TRK)		SMD
STC 8380	3	2500	16 MS (AVG)		IBM
IBM 3380	3	1520	16 MS (AVG)	$116K	IBM
STC 6654	1.2	1270	23 MS (AVG)	$40K	SMD
CENTURY AMS 571	1.92	590	25 MS (AVG)	$10K	SMD
FUJITSU EAGLE	1.86	474	5 MS (1TRK)	$18K	MOD. SMD

OPTICAL DRIVES

STC 7600	3	4000	7 MS (1 TRK)	$130K	IBM
SHUGART OPTIMEM	0.5	1000	100 MS (MAX)	$6000	SCSI
CDC	0.2	1000	260 MS (MAX)	$25K	
RCA "JUKEBOX"	6.25	9.75 GB	300 MS		CUSTOM

PARALLEL TRANSFER DRIVES

IBIS 1400	10.6	1400	2.5 MS (1 TRK)	$65K	CUSTOM
AMPEX PTD9300	10.8	312	6 MS (1 TRK)	$62K	UNIBUS, DR11W

NOTES:
(1) ALL DRIVE CAPACITIES ARE UNFORMATTED.
(2) SMD = STORAGE MODULE DRIVE
(3) SCSI = SMALL COMPUTER SYSTEM INTERFACE

DISK EMULATORS AND RAM CACHES

MODEL	RATE (MBYTE/SEC)	CAPACITY (MBYTE)	
AMPEX MEGASTORE	2	32	
STC 9305	3	96	
CENTENNIAL SSD		128	
DATARAM BS320	64	32	($161k)
CRAY SSD	100	256	

APPENDIX B: IBIS INTERFACE SUMMARY

The IBIS drive uses a sixteen bit data channel to achieve a data transfer rate of 24 MByte/sec (burst). All signals are driven by differential driver/receiver pairs. Drivers are type 75110A, and receivers are type 75108A. Cables are 20 twisted pair (40 wires total), and connectors are standard ribbon cable types. A detailed description of the IBIS interface can be found in the IBIS Systems Inc. Model 1400 Interface Specification.

BUS INTERFACE SIGNALS

PIN	BUS CABLE	CONTROL CABLE
1	GND	GND
2	GND	GND
3	BUS 00+	FUNCTION RDY+
4	BUS 00-	FUNCTION RDY-
5	BUS 01+	READY+
6	BUS 01-	READY-
7	BUS 02+	RDCLK+
8	BUS 02-	RDCLK-
9	BUS 03+	ERROR+
10	BUS 03-	ERROR-
11	BUS 04+	WRCLK+
12	BUS 04-	WRCLK-
13	BUS 05+	SELECTED+
14	BUS 05-	SELECTED-
15	BUS 06+	BUSY+
16	BUS 06-	BUSY-
17	BUS 07+	CODE 0+
18	BUS 07-	CODE 0-
19	BUS 08+	CODE 1+
20	BUS 08-	CODE 1-
21	BUS 09+	CODE 2+
22	BUS 09-	CODE 2-
23	BUS 10+	CODE 3+
24	BUS 10-	CODE 3-
25	BUS 11+	CODE P+
26	BUS 11-	CODE P-
27	BUS 12+	BUS SAFE/
28	BUS 12-	GND
29	BUS 13+	DATA REQ +
30	BUS 13-	DATA REQ -
31	BUS 14+	DIR IN +
32	BUS 14-	DIR IN -
33	BUS 15+	RESET +
34	BUS 15-	RESET -
35	BUS P+	STATUS P+
36	BUS P-	STATUS P-

37	DATA RDY+	DEVICE ENB 0+
38	DATA RDY-	DEVICE ENB 0-
39	RESERVED	DEVICE ENB 1+
40	RESERVED	DEVICE ENB 1-

BUS00 - BUS15 (BIDIRECTIONAL)
These signals form the sixteen-bit data bus between the drive and the controller. The direction of the bus is controlled by DIR IN.

BUS P (BIDIRECTIONAL)
This line forms the odd parity of the data bus. BUS P is valid only on bus cycles during which DATA READY is valid. DIR IN controls the direction of this signal.

DATA RDY (BIDIRECTIONAL)
This signal is asserted by the device driving the bus to indicate that the data on BUS00 - BUS15 is valid. DIR IN controls the direction of this signal.

FUNCTION READY
This signal, driven by the controller, indicates that the data on CODE0 - CODE3 is valid.

READY
READY is driven by the enabled drive to indicate that the drive is ready to accept commands.

RDCLK (READ CLOCK)
This signal is generated by the enabled drive to synchronize data and status presented to the controller. Data and parity are valid 20 ns before and after the falling edge of RDCLK. The period of RDCLK is 99.5 - 100.5 ns.

ERROR
ERROR is valid for at least one RDCLK cycle before and after the trailing edge of BUSY to indicate that an error or drive fault condition has been detected.

WRCLK (WRITE CLOCK)
This signal is generated by the controller to synchronize data and commands to the drive. Data and parity are valid 20 ns before and after the falling edge of WRCLK. The period of WRCLK is 99.5 - 100.5 ns.

SELECTED
This signal is asserted by the drive in response to a valid SELECT or RELEASE OPPOSITE AND SELECT command. SELECTED is

activated 200 ns prior to the trailing edge of BUSY.

BUSY

BUSY is driven by the enabled drive to indicate that the drive is executing a command.

CODE 0 - CODE 3

These lines are driven by the controller to transmit commands to the drives. These lines are valid when FUNCTION READY is valid. Bit 0 is the most significant bit.

FUNCTION CODES

0 - 3	COMMAND
0	ECHO
1	SELECT
2	READ
3	WRITE
4	HEAD SELECT
5	CYLINDER SELECT
6	(UNUSED)
7	SELECT STATUS
8	GENERAL STATUS
9	DIAGNOSTIC
A	(UNUSED)
B	(UNUSED)
C	CLEAR FAULTS
D	RETURN TO ZERO
E	RELEASE OPPOSITE CHANNEL AND SELECT
F	RELEASE

CODE P

This signal gives the odd parity of CODE 0-3.

DATA REQ (BIDIRECTIONAL)

This signal is asserted to request data during write data, write buffer, and read data commands. When DIR IN is asserted, the controller sources this signal, and when DIR IN is not asserted, the drive is the source.

DIR IN

This signal is asserted by the controller to indicate that the enabled drive is driving the bus.

BUS SAFE/

This is an active low, single ended line used by the drive to detect open cable and controller-powered-down conditions. The controller drives this line with a 74S38, or equivalent driver.

STATUS P

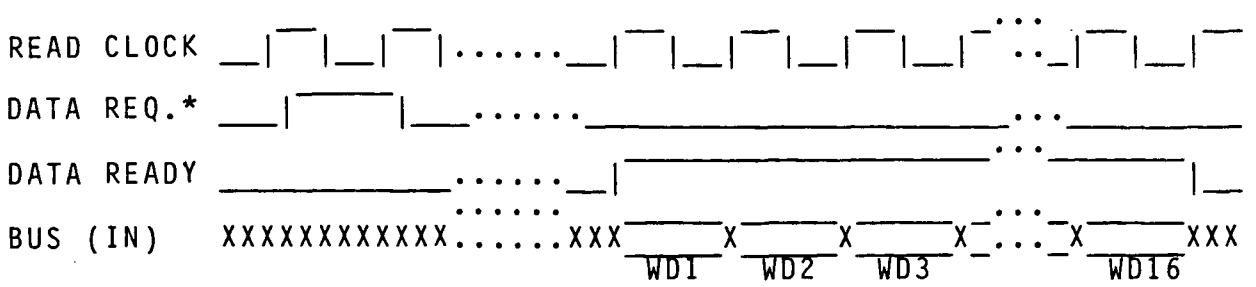
This signal forms the odd parity of READY, ERROR, SELECTED, and BUSY. This signal is driven by the enabled drive. Parity is valid on bus cycles when READY is active.

RESET

This signal is driven by the controller and is used to reset all drives on the bus.

DEVICE ENB 0, 1

These signals are driven by the host to enable one of four possible drives onto the bus. DEVICE ENB 0 is the least significant bit.



* DATA REQ. IS PRESENTED ONLY ON 2ND AND SUBSEQUENT TRANSACTIONS

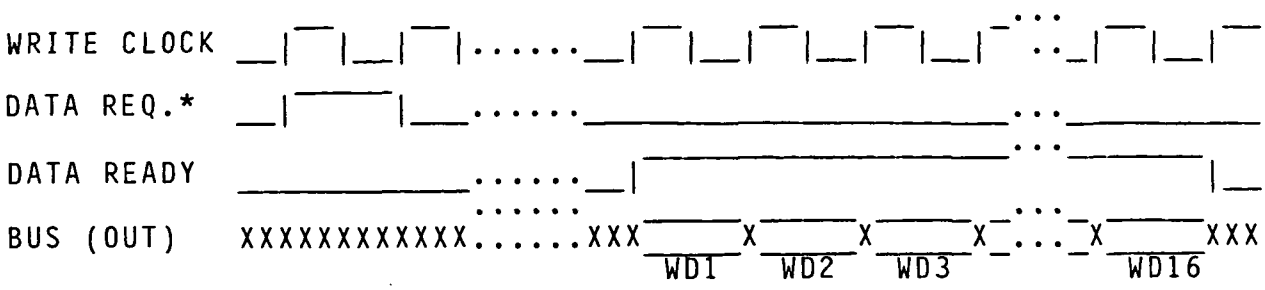
SETUP & HOLD TIME, DATA RELATIVE TO FALLING EDGE OF RDCLK: 20 NS

RDCLK PERIOD IS 100 NS

AVERAGE TIME BETWEEN 16-WORD PACKETS IS 900 NS

DATA RECORDS ARE READ IN 4K-BYTE BLOCKS

FIGURE B-1: READ DATA TIMING



* DATA REQUEST IS PRESENTED ON ALL PACKET TRANSFERS

SETUP & HOLD TIME, DATA RELATIVE TO FALLING EDGE OF WRTCLK: 20 NS

WRTCLK PERIOD IS 100 NS

AVERAGE TIME BETWEEN 16-WORD PACKETS IS 900 NS

DATA RECORDS ARE WRITTEN IN 4K-BYTE BLOCKS

FIGURE B-2: WRITE DATA TIMING

APPENDIX C: ALTERNATIVE SYSTEM

C.1 APTEC - IBIS SYSTEM OVERVIEW

In this appendix an alternate approach to an MPP disk subsystem is discussed. This alternate approach capitalizes on disk system components which have recently become available in the commercial market. In particular, the design employs the Aptec Dimensional Processing System (DPS) which provides hardware and software support for high performance processing and storage devices for DEC systems.

The DPS-2400 system is designed to integrate multiple system components to a VAX or PDP-11 computer, with enhanced mass storage, without loading down the host computer. Aptec has announced an interface to the Ibis 1400 disk drive. A possible system configuration for the MPP with the Aptec system is shown in Fig. C-1.

The DPS-2400 is designed primarily as a way of attaching high-speed peripherals, such as array processors, to VAX series host computers without placing these peripherals on the backplane of the host computer. Aptec does this by connecting the peripherals to a "private bus" which is a high speed version of the DEC Unibus. Aptec uses intelligent Data Interchange Adapters (DIAs) and Data Interchange Processors (DIPs) to control these peripherals, and uses the private bus and a Data Interchange Bus (DIB) as communications pathways between peripherals. The DIAs and DIPs are constructed in such a way that all peripherals can communicate with the host computer as though they were directly connected to the Unibus. Also attached to the DIB are one or more mass memory modules, which may contain data or instructions for the DIAs and DIPs

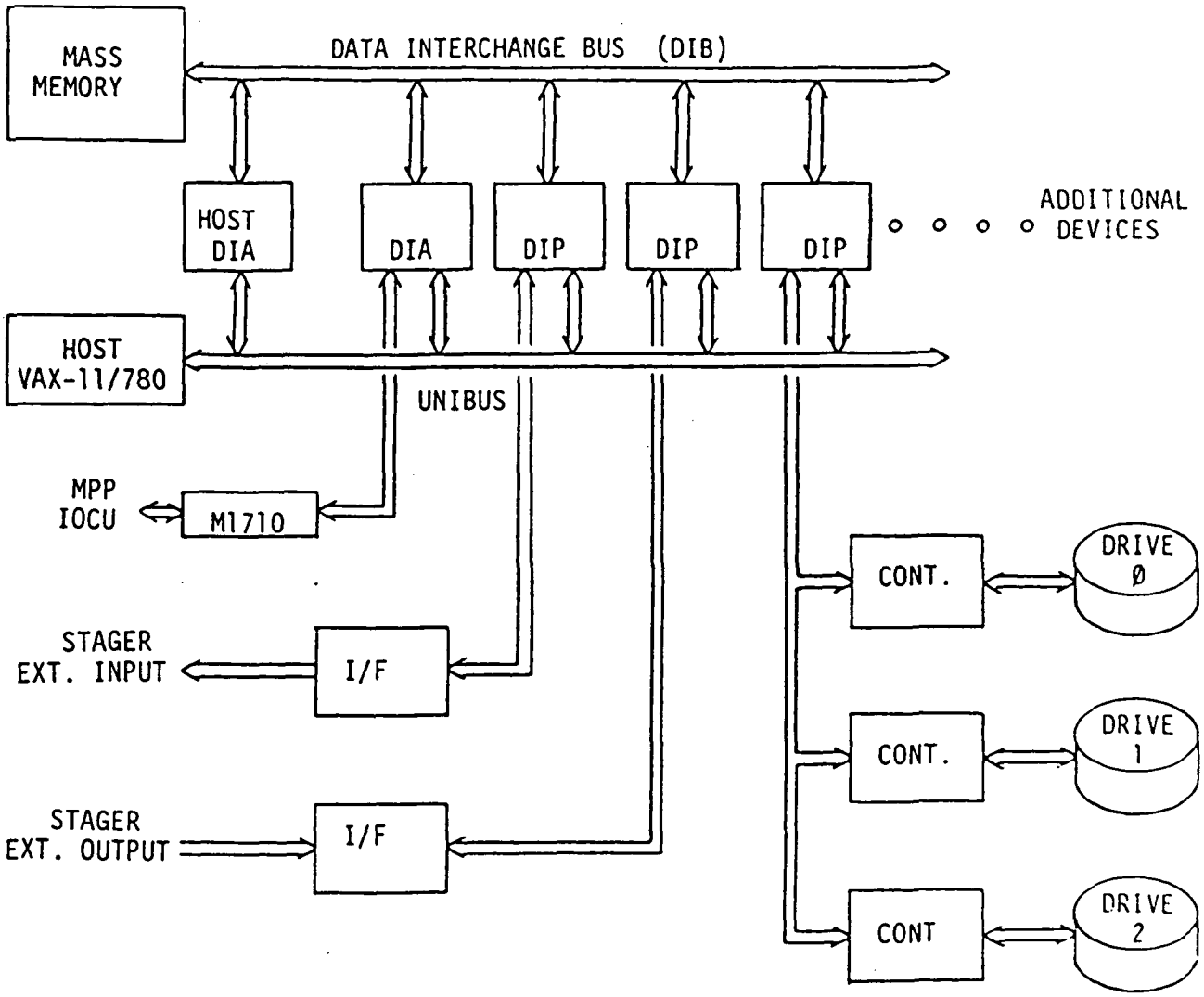


Figure C-1. MPP-APTEC DPS DISK SYSTEM

C.2 DPS-2400 HARDWARE

C.2.1 DATA INTERCHANGE ADAPTER AND DATA INTERCHANGE PROCESSOR

The Data Interchange Adapters (DIA), and the higher-speed Data Interchange Processor (DIP), are bit-slice controlled devices whose microcode is completely downloadable from the host computer. In this way, the DIAs and DIPs can be programmed to be high-speed peripheral controllers, general purpose computers, or DPS-2400 network controllers. Aptec provides the software tools, including a high level language, for programming these devices.

Each DIA or DIP is accessable to the host as a standard Unibus device. One DIA, called the Host DIA (HDIA) is responsible for the primary link between the host computer and the remainder of the system. The remaining devices are avaliable for use as intelligent peripheral controllers or links to additional downstream DPS-2400s.

In the alternate system block diagram, shown in Figure C-1, one DIA is used as a controller to the M1710 which is used by the MPP to present I/O interrupts to the host computer. By removing the M1710 from the VAX Unibus and placing it on the DPS-2400 private bus, these interrupts can be processed by the DIA, and the necessary commands passed on to the drive adapters, without going through the overhead time necesary when interrupting the host computer. However, since the private bus appears to the VAX as an extension of the UNIBUS, the host VAX can still access and control the MPP as well.

One DIP in the alternate system is used to provide a command and data channel to up to three disk adapters. Each disk adapter is capable of controlling one Ibis disk drive.

Two DIPs are allocated to the task of interfacing to the MPP Staging Memory. One DIP receives stager output data, and one provides input data to the stager. Note that since these are intelligent controllers, stager input and output functions can occur simultaneously.

The private bus of the DIP employs a multiplexed, eight-bit command and data bus which supports both word-by-word asynchronous transfers and burst synchronous data transfers. The asynchronous mode supports transfer rates up to 4 MByte/s, while the synchronous mode is used for transfer rates of up to 12 MByte/s.

C.2.2 DPS-2400 MASS MEMORY

The DPS-2400 is capable of accessing up to 4GBytes of Mass Memory. The Mass Memory cards contain 1 MByte of Dynamic RAM each, organized into 32-bit words, with 12-bit error correction/detection fields. It is anticipated that approximately 4 Mass Memory cards would be required for the

alternate system.

C.2.3 IBIS DISK ADAPTER

The Aptec Ibis Disk Adapter consists of a single hex PCB which plugs into any available slot in the DPS-2400 backplane. It is capable of controlling one Ibis model 1400 disk drive. The Ibis Disk Adapter enables any other node in the DPS-2400 system, including the host computer, to access files on the disk drive using the FILES-11 file system. The Ibis Disk Adapter connects to the remainder of the DPS-2400 through a DIP. Each DIP can support up to three Ibis Disk Adapters.

C.2.4 STAGER INTERFACE HARDWARE

The MPP Staging Memory communicates to the DPS-2400 through two Stager Interface cards. These two cards represent the only new designs in the alternate system. (The redesign of the Stager I/O boards would still be necessary, though.) The Stager Interface cards are responsible for translating protocols between the stager I/O protocol in section 4 to the DIP I/O protocol described below.

C.3 SOFTWARE

Aptec provides a software library with the DPS-2400 system which includes the drivers necessary to run the DIAs, DIPs, and Ibis adapters. Aptec also provides a software development package, including a high-level language, for implementing custom interfaces to peripheral devices. It will be necessary to write new software to drive the stager interface hardware. Since the existing software supports the FILES-11 data structure of the VAX, new software for the Ibis adapters will not need to be written.

C.4 EXPANSION

The alternate system can be implemented with one row of up to three disk drives. It is also possible to expand this system to multiple rows of drives by using additional DPS-2400 systems, daisy-chained off the initial DPS-2400 through the private bus of a DIP. In the expanded system, each row of drives would buffer data in the mass memory of its DPS-2400. The data would be read or written to the stager through the Stager Interface boards. In the expanded alternate system, the mass memory would serve the same purpose as the data buffers in the proposed system: that of deskewing the data between the rows of disk drives. The stager interface boards would communicate with each other in order to present data simultaneously to the stager.

This expanded system would require modifications to the existing file handling software, since data files would now be split up among multiple disk drives, which removes the data structure from the FILES-11 convention.

C.5 PHYSICAL

The DPS-2400 system fits in a standard 19 inch RETMA cabinet. The overall dimensions of the cabinet are 24 inches wide by 30 inches deep by 80 inches high. The DPS-2400 contains integral power supplies which run off of 120 VAC. The stager interface cards would be located in the same cabinet as the DPS-2400 system(s). The subsystem chassis would be located adjacent to the MPP chassis.

C.6 COST

The cost projection below is based on present system configuration knowlege and understanding. It is presented for budgetary purposes only and should not be interpreted as a price quote for the project.

ITEM	COST
1. MANAGEMENT	$100K
2. HARDWARE ENGINEERING	$150K
3. SOFTWARE ENGINEERING	$ 50K
4. APTEC HARDWARE *	
MAIN CHASSIS INCL. 2 DIAs, 1MBYTE MEMORY, POWER SUPPLIES, SOFTWARE, DIB CONTROLLER	$ 32.5K
3 DIPs	$ 27K
2 MBYTE ADDITIONAL MEMORY	$ 14K
2 IBIS CONTROLLERS	$ 60K
5. IBIS DISK DRIVES (2)	$134K
6. QA, MANUFACTURING	$ 15K
7. CHECKOUT	$100K
TOTAL FACTORY COST	$682.5K
SELL PRICE	$871K

* Price based on verbal price quotes from Aptec, no formal quotes obtained due to the development nature of the Aptec system components.

C.7 CONCLUSIONS

The Aptec system holds promise as an inexpensive alternate to the proposed system. It utilizes hardware which is mostly available off the shelf, and most of the software for the system is also available.

Certain critical components to the Aptec system, however, are not currently available. The DIP is currently in the prototype stage, and the Ibis disk adapter is not yet completely designed.